哈爾濱工業大學（深圳）
HARBIN INSTITUTE OF TECHNOLOGY

实验报告

开课学期：2020 秋季

课程名称：大数据导论

实验名称：数据理解、数据预处理及决策树的应用

实验性质：设计型

实验学时：2 地点：T2608

学生班级：2018 级计算机 5 班

学生学号：180110505

学生姓名：胡聪

评阅教师：

报告成绩：

实验与创新实践教育中心制

2020 年 7 月

一、 实验目的

1. 学会理解数据并对数据进行预处理；

2. 理解决策树的原理并掌握其构建方法。

二、 实验内容

1. 熟悉 Pandas 的安装和使用，并对数据进行预处理和可视化分析；

2. 熟悉 sklearn 包，调用决策树模型对数据进行训练。

三、 实验过程

1. 启动 jupyter

2. 安装 Pandas 库并熟悉其基本操作

（1）生成数据

（2）计算数据的基本信息

（3）选取特定列

（4）选取特定列行

（5）选取多行多列

（6）选取 B 列大于等于 5 的数据

（7）修改指定列

3. 数据读取及预处理

4. 安装 sklearn 并构建决策树

5. 参数调整

6. 可视化决策树模型

参数调整过程：

（1）criterion：用以设置用信息熵还是基尼系数计算

基于基尼系数进行计算

```
dtc = DTC(criterion='gini',max_depth=5) #基于基尼系数
dtc.fit(X_train,y_train)
print('准确率',dtc.score(X_test,y_test))
```

```
准确率 0.9001919385796545
```

基于信息熵进行计算

```
dtc = DTC(criterion='entropy',max_depth=5) #基于信息熵
dtc.fit(X_train,y_train)
print('准确率',dtc.score(X_test,y_test))
```

```
准确率 0.8886756238003839
```

可以看到采用基尼系数进行计算，准确率更高

（2）splitter：指定分支模式

默认模式为 best，表示选择最优的分裂策略

```
dtc = DTC(criterion='gini',max_depth=5) #基于基尼系数
dtc.fit(X_train,y_train)
print('准确率',dtc.score(X_test,y_test))
```

```
准确率 0.9001919385796545
```

改为 random，表示选择最好的随机切分策略，发现准确率发生了下降

```
[29]: dtc = DTC(criterion='gini',max_depth=5,splitter='random') #基于基尼系数
      dtc.fit(X_train,y_train)
      print('准确率',dtc.score(X_test,y_test))
```

```
准确率 0.8886756238003839
```

（3）max_depth：最大深度，防止过拟合

```
for depth in range(1,10):
    dtc = DTC(criterion='gini',max_depth=depth) #基于基尼系数
    dtc.fit(X_train,y_train)
    print('depth:',depth,'|','准确率',dtc.score(X_test,y_test))
```

```
depth: 1 | 准确率 0.8714011516314779
depth: 2 | 准确率 0.8925143953934741
depth: 3 | 准确率 0.8944337811900192
depth: 4 | 准确率 0.8982725527831094
depth: 5 | 准确率 0.9001919385796545
depth: 6 | 准确率 0.8925143953934741
depth: 7 | 准确率 0.8771593090211133
depth: 8 | 准确率 0.8656429942418427
depth: 9 | 准确率 0.8541266794625719
```

可以看到在 depth 为 5 的时候准确率最高

（4）min_samples_leaf：限定每个节点分枝后子节点至少有多少个数据，否则就不分枝

整数：

```
dtc = DTC(criterion='gini',max_depth=5,min_samples_leaf=1) #基于基尼系数
dtc.fit(X_train,y_train)
print('准确率',dtc.score(X_test,y_test))
```

```
准确率 0.9001919385796545
```

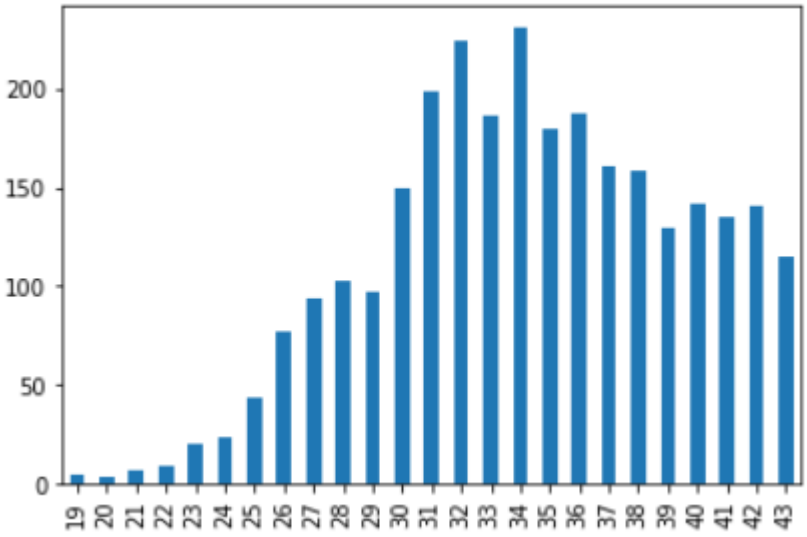
浮点数：

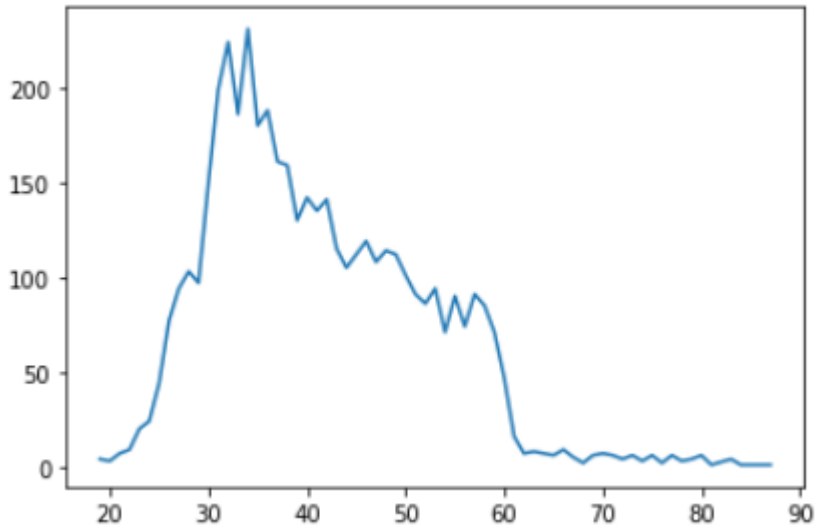
```
dtc = DTC(criterion='gini',max_depth=5,min_samples_leaf=0.1) #基于基尼系数
dtc.fit(X_train,y_train)
print('准确率',dtc.score(X_test,y_test))
```

```
准确率 0.8790786948176583
```

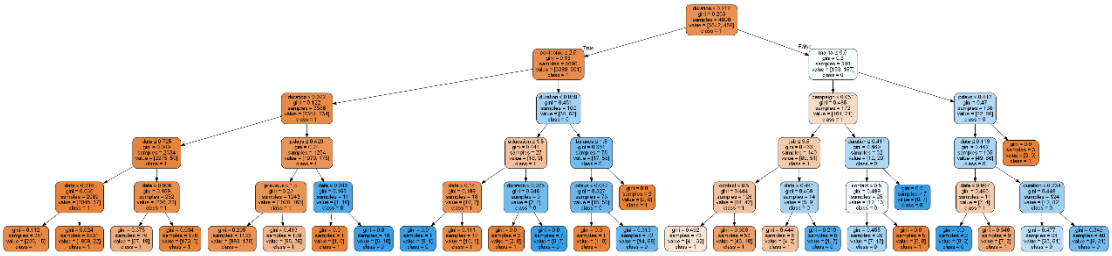
四、 实验结果与分析

数据可视化


200
150
100
50
0
19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43


200
150
100
50
0
20 30 40 50 60 70 80 90

可视化决策树模型



个人签名 ：

2020 年　　12 月　　10 日